The physiological mechanisms concerned in the development of neuroses have been fairly fully investigated by the Pavlov school.

The Freudian method of psycho-analysis, employed for the study of personality, is tendentious, as it amounts merely to a sexual interpretation. The attempts by some investigators, particularly in the U.S.A. to combine Freudism with the Pavlov doctrine are, from our point of view both fruitless and unnecessary. The attitude of Pavlov himself to Freud was generally one of negation, although he took some material described by Freud to illustrate his views.

Progress Report No. 2 Jan. 1963 COMPUTER-AIDED RESEARCH IN MACHINE TRANSLATION

PHYSIOLOGICAL MECHANISMS OF THE RISING OF NEUROSES
WERE SUFFICIENTLY (RATHER) TOTALLY (FULLY) STUDIED BY PAVLOV (PAVLOVIAN) SCHOOL.

THE FREUDIAN METHOD OF THE PSYCHOANALYSIS, USED FOR THE STUDY (INVESTIGATION) OF THE PERSONALITY (PERSON) IS TENDENTIOUS, (BIASED) SINCE IT IS REDUCED TO THE SEXUAL INTERPRETATION. ATTEMPTS OF CERTAIN (SOME) INVESTIGATORS (RESEARCHERS), IN PARTICULAR IN THE USA, TO UNITE (INTO) FREJDIZM WITH THE DOCTRINE OF PAVLOV, IS, FROM OUR POINT OF THE VISION, (SIGHT, EYESIGHT, VIEW) FRUITLESS AND UNNECESSARY.

(9)

PROGRESS REPORT NO. 2

Under Contract NSF-C233

With National Science Foundation

(6) COMPUTER-AIDED RESEARCH IN MACHINE TRANSLATION

(7) NH

(8) 21

(10) NH

(11) 31 January 1963

(12) N.

(13) NH

(15 Time 17) NH

(20) 11

(14) Rpt. no. C157-3U1

(5) THOMPSON RAMO WOOLDRIDGE INC.
RW DIVISION
CANOGA PARK, CALIFORNIA

(35) 873 050

PREFACE

The Machine Translation Group at Thompson Ramo Wooldridge Inc. has been receiving the support of two different sponsoring agencies. The work covered in the present report was done under a cost-sharing contract, NSF-C233, with the National Science Foundation.

Our research benefited from the technical interest shown by Richard See, Assistant Program Director for Mechanical Translation, Documentation Research Program.

Some of the tools used in this research had been created under previous contracts with the Intelligence Laboratory, Rome Air Development Center, Griffis Air Force Base. Many of the tools created under the present NSF contract will be used to further additional research now being sponsored by RADC. In general, work being done at the RW Division to improve the techniques for research in machine translation has been done under contract to the NSF, while studies in the area of semantics have been done under contract with RADC.

The support of the National Science Foundation is hereby gratefully acknowledged.

The research on this project was primarily performed by:

Jules Mersel (Principal Investigator)
Gerhard Reitz (Associate Project Manager)
Paul L. Garvin
Herbert H. Holley
Christine A. Montgomery
George Onischenko
Steven B. Smith

CONTENTS

PART I

THE TRW TRANSLATION ERROR DETECTOR

(TED)

SMALL OUTGROWTHS IN THE [FORM] OF CONES OR PROTUBERANCES OF [THE] LARGER DIMENSION [WERE] FORMED IN OTHER CASES.

This sentence is the machine translation of a Russian sentence taken from the Russian Journal Eksperimental'naya Morfologiya.
What is wrong with this translation?

If we compare it to an independent human translation of the same Russian sentence, we find that the boxed word "the" is most conspicuously wrong. It is one of three words shown in boxes in the above sample sentence. The three boxed words were identified by a separate computer program to be different in the machine translation from an independent human translation. How did the computer identify this difference between machine translation and human translation? Let us go back to the original Russian sentence which appeared in an article entitled "Restoration of the Regenerative Capacity of the Extremities in Axolotls Depressed by Roentgen Radiation," by L. V. Polezhaev and N. I. Ermakova in the above named journal. It is shown in Figure 1. The original Russian sentence appears on the top line, the middle line shows its transliteration into Roman letters, the bottom line shows its word-for-word machine translation.

В	других	случаях	образовывались	небольшие	выросты	
V	DRUGIX	SLU(A=X	OBRAZOVYVALIS*	NEBOL*)IE	VYROSTY	
IN	OTHER	CASES OCCASIONS EVENTS INCIDENTS	WERE	PRODUCED FORM EDUCATE	SMALL	PROTUBERANCES OUTGROWTHS

в	виде	конусов	или	выросты	большего	размера
V	VIDE	KONUSOV	ILI	VYROSTY	BOL*)EGO	RAZMERA.
IN	THE FORM	OF CONES	OR	PROTUBERANCES OUTGROWTHS	OF THE LARGER	DIMENSION.

Figure 1.

The underlined words on the third, English, line have been inserted by the translation program because, though nonexistent in the Russian, they are necessary for the English translation. These insertions are based on a thorough automatic parsing of the original Russian. Nevertheless, the insertions are often wrong. In the case of article insertion, the situation is most difficult to remedy, since — as is well known — there are no articles in the Russian language. But even in the case of prepositions and auxiliary verbs, for which definite clues are present in Russian, the program makes many insertion "mistakes".

The first step in the correction of these translation errors is their identification. The only way of doing it, until recently, has been to employ skilled personnel to "post-edit" the machine translation and keep a record of the mistakes, or to compare the machine translation to an existing human translation and record the differences. Both methods are laborious, and personnel who are skilled enough to do the job correctly are not willing to subject themselves to the necessary tedium, and conversely. The optimal solution to the problem of translation error detection, therefore, is automation.

We have developed a computer program which will match a professional human translation sentence for sentence with a machine translation, then compare the two translations word for word. The only human labor required for this automatic comparison is keypunching.

Figure 2 schematizes the comparison of the human and machine translations of the sample sentence cited in the beginning of this section. The human translation is shown in the right-hand column, the machine translation in the left-hand column. In the Russian original, the verb (OBRAZOVYVALIS* "were produced") headed the sentence. To make the translation conform to English syntax, the original word order has been rearranged as shown in Figure 2. Similarly, the multiple equivalents produced by the machine translation program have been reduced to one matching translation each by the matching program. Thus, the human translation "formed" found a match with one of the three equivalents PRODUCED. The matching program was able to associate the machine
FORM
EDUCATE

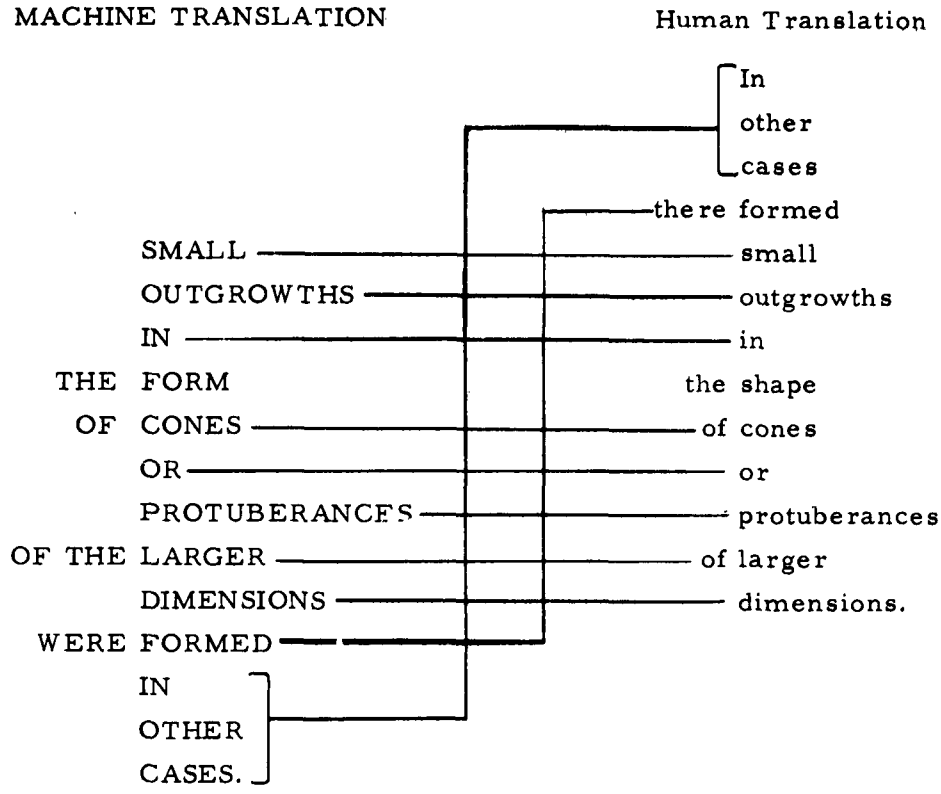

MACHINE TRANSLATION
Human Translation
IN OTHER CASES.
SMALL OUTGROWTHS IN THE FORM OF CONES OR PROTUBERANCES OF THE LARGER DIMENSIONS WERE FORMED
In other cases
there formed
small outgrowths in the shape of cones or protuberances of larger dimensions.

Figure 2.

translation equivalent FORM with the word "formed" in the human translation because it has the capability for recognizing grammatical endings such as "-ed" and hence for matching different grammatical forms of the same word.

```
  QB  8    6   12   201000  2   VIDE                    THE FORM
      THERE                 . .   4
      SHAPE                 . .   10
225RS 225ES  10  MA  11  TO  90  QU  0  SW  0  DF  0  MISM
```

Figure 3.

As we can see from Figure 3, the printed output of the matching program, there were two words in the human translation which did not match, namely, the fourth and tenth words ("there", and "shape"). On the other hand, the Russian word **виде**, in our transcription VIDE, translated as THE FORM, was left unmatched. A linguistic analyst can infer from this printout that the Russian word VIDE, which the machine translated as THE FORM, was rendered by SHAPE in the human translation. Note how useful this type of information is for updating a Russian-English dictionary; in this case add "shape" to the translations of Russian "VID".

Our output, however, not only indicates the words that did not find their matches in the matching program. It also automatically compiles a running count of the number of words that have found their matches. This is shown on the last line of Figure 3 which indicates that ten words matched out of a total of a total of eleven, giving a matching quotient of 90%.

The main purpose of this matching quotient is to determine automatically whether the sentence of the Russian original and the machine translation corresponded to that of the human translation or whether the sentences were "out of step". Say, one sentence of the original corresponding to two sentences in the human translation, or conversely. The high matching quotient is a sure sign that the sentences corresponded to each other.

A significant feature of our machine translation capability is the insertion into the English translation of function words (such as articles, prepositions, verbal auxiliaries) that are not present in the Russian original but are required for the English. One important purpose of the matching program is to determine whether these insertions are adequate. The program does this by comparing the insertion of a number of function words in the machine translation with the words that correspond to them in the human translation. The correspondence is established by recording a matched word adjacent to which there has been an insertion in the machine translation (e.g., insertion of "OF THE" before "LARGER" as shown on Figure 2), and checking the word on the same side of the match of this word in the human translation (in our example, "OF" before "LARGER" in the human translation as shown in Figure 2). In this example, we correctly inserted the preposition "OF", corresponding to the Russian genitive BOL*)EGO, but the additional insertion of the article "THE" was incorrect.

A statistical record is kept by the matching program of the correspondence of the insertion of a number of function words in the machine translation and the human translation. The function words of interest are the three written article forms (the, a, an), 11 prepositions, and 25 verbal auxiliaries. For each of these words, we reserve four counters in the computer: (a) the same insertion has been made in the machine translation as in the human translation (called "corresponding insertion"); (b) an insertion has been made in both the machine translation and the human translation, but the inserted words are not the same (called "non-corresponding insertion"); (c) no insertion has been made in either the machine translation or the human translation (called "corresponding non-insertion"); (d) an insertion has been made in the machine translation where none has been made in the human translation, or conversely, no insertion has been made in the machine translation where one has been made in the human translation (non-corresponding non-insertion). The use of the counters is illustrated in Figure 4 below:

Counter	exemplified by Machine translation having	Where human translation has
(a) Corresponding insertion	have	have
(b) Non-corresponding insertion	a	the
(c) Corresponding non-insertion	---	---
(d) Non-corresponding non-insertion	were	----
	---	were

Figure 4.

The printout of the statistics of the correspondence of function-word insertions in the sentence found in Figure 2 is shown in Figure 5 below.

	Ctr(a)	Ctr(b)	Ctr(c)	Ctr(d)
THE			3	1
OF	2			
IN	1			
WERE				1

Figure 5.

A statistically significant score requires data from more than one sentence—all the sentences of a fair sized article may be sufficient. The statistics for the biology article from which the sentence found in Figure 2 was taken are shown in Figure 6.

	(a)	(b)	(c)	(d)		(a)	(b)	(c)	(d)		(a)	(b)	(c)	(d)
THE	100	125	227	70	A	0	14	0	2	OF	99	64	0	6
IN	20	18	0	7	AT	0	14	0	0	TO	2	0	0	1
FOR	1	1	0	1	AFTER	22	0	0	1	BY	3	8	0	2
FROM	3	1	0	0	WITH	5	25	0	0	ON	7	0	0	0
IS	1	2	0	0	ARE	1	2	0	1	WAS	4	7	0	2
DID	3	0	0	0	CAN	1	0	0	0	BE	0	1	0	0
WERE	4	4	0	13										

Figure 6.

A quick glance shows that only 2 of 3 article forms were used in MT or human translation, only 9 out of 11 prepositions, and 7 out of 25 auxiliaries. "THE" was inserted correctly 100 times, left out correctly 227 times, was inserted incorrectly in the MT 125 times, and left out incorrectly 70 times. There were 327 correct occurrences out of a total of 552, or 62.6 percent.

To the 16 cases in which "A" appeared in the human translation corresponded the incorrect insertion of "THE" in the machine translation in 14 cases.

Our best score among the prepositions was achieved with "AFTER", which we used correctly 22 out of 23 times, and "ON", which was correctly used 7 out of 7 times. Our worst was predictably with "AT", which was consistently misused.

Our statistics of auxiliary insertion are inadequate. The only auxiliary for which we had meaningful statistics was "WERE"; this auxiliary will require a great deal of attention, since we only inserted incorrectly in about 15 percent of the cases.

A detailed discussion of the matching and checking program is found in Section 18 of Part II. The flowcharts appear in Appendix B.

The statistical tabulation of correct and incorrect insertions (and non-insertions) is, however, not enough. In order to improve the machine translation program, the linguists will want to study the "offending" sentences — that is, those sentences in which the non-corresponding counters (b) and (d) show too large a percentage. We have written a program which retrieves these problem sentences; it is facetiously called "bôche-fèchre" (pronounced botch-fetcher).

As a result of this study the linguist will formulate new or revised translation or insertion rules. These will be included in the machine translation program and checked out by the programmer. The matching program will then compare the machine translation and human translation and compile statistics in the operation of the new or revised rules. The feedback cycle is now completed.

One more cycle was run for the insertion of the article "THE". After looking at the original statistics, we decided to change the rule and eliminate the insertion of "THE" in front of a genitive nominal block starting with an adjective.

Ten articles in the field of biology were rerun using this new rule. The old and new statistics for the article "Cyto-histological Characteristics of Reparative Processes in Castrates of Various Ages Subsequent to the Administration of Cortisone and ACTH," by A. I. Bukhonova are shown in Figure 7 below:

		Counter			
		(a)	(b)	(c)	(d)
"THE"	Old Way	106	82	123	59
	New Way	104	81	124	61
"A"	Old Way	0	17	0	2
	New Way	0	17	0	2

Figure 7.

As can be seen, very few changes resulted from this change in rule. One more correct omission of "THE" than with the old rule was recorded, with a corresponding decrease in the category of noncorresponding insertions; on the other hand, we lowered our successful matches by two, with a corresponding increase of wrong insertions of "THE". We were able to ascertain the triviality of this rule change without extensive post-editing.

PART II

TECHNICAL DISCUSSION OF WORK PERFORMED

Summary of Work Performed

The following tasks have been accomplished during the reporting period. These are listed below going from input to output:

(1) Programs in constant use have been changed over to the programming system at Space Technology Laboratories.

(2) The Russian and English keypunching instructions have been revised.

(3) The Russian and English edit routines have been revised.

(4) Seven different fields of study were chosen: Biology, Botany, Education, Fiction, Pavlovian Psychology, Soil Science and Cybernetics. The selected Russian and English text has been keypunched.

(5) All Russian and English text has been edited.

(6) A program for listing words missing in the dictionary (new words) has been written and checked out.

(7) The new words from the different text fields have been selected for inclusion in the machine glossary and grammar-coded.

(8) The machine dictionary has been updated with the new words.

(9) A new dictionary printout format has been devised and put into practice.

(10) The entire dictionary has been sorted on the grammar code.

(11) A dictionary duplication discriminator program has been written and checked out.

(12) Statistics of the word-for-word dictionary lookup have been compiled.

(13) New flowcharts have been drawn up for the syntax program.

(14) Some changes have been introduced into the syntax program to adapt it to the Translation Error Detector.

(15) Fail-safe features have been included in the syntax program.

(16) The greater part of the text in the chosen fields has been machine-translated.

(17) A routine to produce a concordance of the translated text has been programmed.

(18) The rules for the Translation Error Detector (TED) have been programmed and checked out, and the program has been applied to several fields.

(19) A program to retrieve translated sentences according to the errors detected (the "bôche-fèchre") has been coded.

(20) A survey of the area of Chinese-English machine translation was undertaken.

1. Change-over to Programming System at Space Technology Laboratories

The following programs have been changed to Space Technology Laboratories' programming system: English Edit, Russian Edit, Dictionary Lookup, Dictionary Print, Stem Affixing and Reinflection, Sentence Reforming, Syntax, Dictionary Revision. The change-over provides the following important advantages for the programs concerned:

a. System B runs (that is, runs of programs that have been changed over to the STL system) can be executed during the day, which allows 2-3 runs a day. Non-system runs can be executed only at night, which limits us to only one run a night.

b. In system B runs, corrections to the program are allowed in symbolic language, i.e., in the same language as programs are written in. This is a vast improvement over the necessity in non-system runs to correct programs in machine language (octal).

c. Programming is facilitated by the use of more than 40 system macros and many programmer-defined macros. A macro is an abbreviation for a block of prototype instruction which, when "filled out", will act as an open subroutine.

d. The debugging of programs, i.e., checking out and testing, is greatly facilitated by special routines built into the STL system which allow printing out of temporary results, tracing of programming loops, dumping of all or parts of the storage areas, etc.

e. In system B, a set of independent programs can easily be connected together in any possible configuration by specially provided operations. This would otherwise be a fairly complex programming job.

2. Revision of Keypunching Instructions

A special transliteration system for keypunching purposes has been introduced, which has served to speed up keypunching of Russian text, to decrease the error rate and to decrease the training period of the operators. This system uses those English letters which visually most resemble the Cyrillic characters (e.g., "A" for "А"; "R" for "Я"; "N" for "И", etc.).

The keypunching code is automatically converted into the linguistically oriented transliteration code which appears in the output. This code resembles conventional transliteration systems based on phonetic equivalence, but with the important difference that it is a one-for-one code, i.e., each Cyrillic character is transliterated by one English letter to ease the problem of outputting. The difference is more conspicuous in the case of Cyrillic characters which are conventionally transliterated by more than one English letter. Thus, "Щ" is transliterated as "W", not as "shch" as is, for instance, recommended by the American Association for the Advancement of Science (7/4/61). See Appendix A.

3. Revision of Edit Routines

The Russian edit routine is being revised in order to incorporate the transliteration changes discussed above.

In addition, the English edit routine has been changed in regard to numerals, Greek letters, and simple equations. Previously, only a record was keypunched of the presence of some unspecified number or symbol string in the original text. The revised routine calls for the

keypunching of the particular numerical, Greek letter, or simple equation (see Appendix B).

4. Selection of Fields of Study

The keypunched text was chosen from the fields of biology, botany, soil science, fiction, education, Pavlovian psychology, and cybernetics. Except for the cybernetics text which was punched in Russian only, at the request of the National Science Foundation, both the Russian text and the corresponding professional English translations were keypunched. The breakdown as to the approximate number of words punched per field is as follows:

	Russian	English
Biology	39,200	48,800
Botany	65,600	79,200
Education	69,600	96,000
Fiction	8,800	9,600
Pavlovian Psychology	43,200	52,000
Soil Science	45,600	62,400
Cybernetics	38,000	
Totals	310,000	348,000

The text in biology included selected articles from Doklady Akademii Nauk for the year 1960. The English translation was published by the American Institute of Biological Sciences. The botany text was selected from the Doklady Akademii Nauk, 1960, botanical sciences sections; translation published by the American Institute of Biological Sciences. The articles on education were taken from Sovetskaya Pedagogika for the year 1959, and the English translation was published by International Sciences Press. The fiction was selected from various sources, including the following: two chapters from War and Peace (Tolstoy), The Station Master (Pushkin), and The Nose (Gogol). The articles on Pavlovian Psychology were taken from the Zhurnal Vysshei Nervnoi Deyatel 'nosti

Imeni I. P. Pavlova for the year 1959. The translations used were published from Pochvovedenie for 1960, translation published by American Institute of Biological Sciences. The cybernetics text punched was a popular Soviet book on the subject, Mashina i Mysl', furnished to us by the National Science Foundation.

5. Editing of Russian and English Text

An example of edited English text appears in Figure 8.

6. New Word Lister Program

Since the operation of the Translation Error Detector would be greatly facilitated if all text words could be found in the machine dictionary, it was considered useful to keep an automatic tally of all words missing in the dictionary, in order to speed up the updating of the dictionary by the inclusion of missing words.

A special program for listing words missing in the dictionary has therefore been written and checked out. This program, called the New Word Lister, provides an alphabetical list of all word forms in a key-punched text that are not contained in our machine glossary and that our stem-affixing procedure cannot identify as being another form of a word present in the glossary. If more than one form of a missing word appears in a text, all the missing forms are listed, since stem-ending analysis is not possible unless at least one form of the paradigm is in the dictionary. Each form is listed in the printout only once.

The New Word Lister also provides us with statistics on the number of new forms encountered in a field not previously processed, and records one text location for an occurrence of a missing form. This record makes it possible to determine in doubtful cases whether a keypunching error has occurred.

For words which cannot be found in the available dictionaries, the record provides a context which may help us to determine its meaning.

PB	TEXT	PA1	PA2	P	S	W	P	L	W	PAGE
	IN			03	02	019	A	29	07	
	LIGHT			03	02	020	A	30	01	
	WERE			03	02	021	A	30	02	
	DISCUSSED			03	02	022	A	30	03	
	EVEN			03	02	023	A	30	04	
	MORE	.		03	02	024	A	30	05	
PARAGRAPH										
*	SPECIAL			04	01	001	A	31	01	
	CAREFULLY			04	01	002	A	31	02	
	REASONED			04	01	003	A	31	03	
	INVESTIGATIONS			04	01	004	A	31	04	
	WERE			04	01	005	A	31	05	
	CARRIED			04	01	006	A	31	06	
	OUT			04	01	007	A	32	01	
	WITH			04	01	008	A	32	02	
	RADIOACTIVE			04	01	009	A	32	03	
	CARBON			04	01	010	A	32	04	
	BY			04	01	011	A	32	05	
*	WEIGL	,		04	01	012	A	32	06	
*	WARRINGTON			04	01	013	A	32	07	
	AND			04	01	014	A	32	08	
*	CALVIN			04	01	015	A	33	01	
(	2	,	)	04	01	016	A	33	02	
*	STEEMAN	-		04	01	017	A	33	03	
*	NIELSEN			04	01	018	A	33	04	
(	3		)	04	01	019	A	33	05	
	AND			04	01	020	A	33	06	
	WITH			04	01	021	A	33	07	
	A	.		04	01	022	A	33	08	
	HEAVY			04	01	023	A	33	09	
	ISOTOPE			04	01	024	A	33	10	
	OF			04	01	025	A	34	01	
	OXYGEN			04	01	026	A	34	02	
*	O			04	01	027	A	34	03	
	BY			04	01	028	A	34	04	
*	BROWN	.		04	01	029	A	34	05	

Figure 8.

A sample page of the output of the New Word Lister is found in Figure 9. The first letter to the right of the Russian word indicates the article in which the word occurred. The second letter indicates the page of the article (the first page of an article is A, the second B, etc.). The first group of numbers indicates the line on the page and the second number the word on the line. For example, SELEKCII occurred in article B, page D, line 11, word 1.

7. Selection and Grammar-Coding of New Words

The output of the New Word Lister serves as a source of new words for addition to the machine dictionary. In the selection of new words for grammar-coding and addition to the dictionary, the following conventions were observed:

In most cases only one form from a given paradigm was selected, since the stem-affixing routine is capable of analyzing all other forms of a regular paradigm on the basis of the presence of one form in the dictionary. In the cases of aberrant Russian paradigms, of English re-inflections which constitute exceptions to the general rules, and in certain other instances, more than one member, and in some cases all the members, of the paradigm had to be selected for the machine glossary.

8. Updating of Machine Dictionary

So far, 6,884 new forms have been added to the machine glossary in seven separate updating runs, giving us coverage increased to approximately 50,000 forms.

An example of large-scale dictionary updating occurred in connection with the cybernetics text.

During the course of the Contract we were requested to translate a 38,000-word book on cybernetics. Since no work had ever been done at the RW Division in this field, we anticipated a sizeable number of missing words. We therefore ran a new word listing. Figure 10 shows the statistics by batches of 4092 words for this text. This figure is explained

SVERXKOMPLEKTNYMI	TC	34	6
SVERXKOMPLEKTNYX	TC	19	5
SVOBODNOSTVUAIYI	EA	3	1
SVOEOBRAZIE	YA	5	7
SV=ZKOJ	RC	2	8
SGLA$IVANIE	SA	28	9
SDVIGANI=	RB	36	4
SEVERNEE	PB	16	6
SEVERNOSO	DA	5	7
SEVERNOJ	HB	42	1
SEVERNYX	HC	36	7
SEGMENTO=OBRNYX	WA	25	6
SEGOLETOK	TA	16	5
SEDALIWNYJ	VA	23	3
SELA	YB	3	2
SELEZENKE	LA	16	8
SELEZENKI	VB	41	1
SELEZENKU	VA	31	3
SELEKCII	BD	11	1
SEMEJ	BD	42	9
SEMENA	GA	38	4
SEMENAMI	GA	13	6
SEMENNIKAX	BC	40	1
SEMENNIKOV	QD	29	6
SEMENNUQ	BD	9	10
SENOKOSY	GA	33	2
SERAXSA	PB	13	10
SERI'N	FD	10	7
SE=ATORNUGGC$	YA	9	12
SZADI	RB	3	2

Figure 9. Sample Page of New Word Lister Output

GLOSSARY LOOKUP
EDITED TEXT CORPUS L CYBERNET

```
IN THIS BATCH
      1007 WORDS WERE FOUND  DIRECTLY ON THE TAPE
       410 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
       324 WORDS WERE FOUND  TO BE MISSING
      2351 WORDS WERE FOUND  TO BE DUPLICATES

IN THIS BATCH
      1072 WORDS WERE FOUND  DIRECTLY ON THE TAPE
       390 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
       286 WORDS WERE FOUND  TO BE MISSING
      2344 WORDS WERE FOUND  TO BE DUPLICATES

IN THIS BATCH
       900 WORDS WERE FOUND  DIRECTLY ON THE TAPE
       326 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
       280 WORDS WERE FOUND  TO BE MISSING
      2586 WORDS WERE FOUND  TO BE DUPLICATES

IN THIS BATCH
      1093 WORDS WERE FOUND  DIRECTLY ON THE TAPE
       466 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
       329 WORDS WERE FOUND  TO BE MISSING
      2204 WORDS WERE FOUND  TO BE DUPLICATES

IN THIS BATCH
      1023 WORDS WERE FOUND  DIRECTLY ON THE TAPE
       333 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
       220 WORDS WERE FOUND  TO BE MISSING
      2516 WORDS WERE FOUND  TO BE DUPLICATES

IN THIS BATCH
      1075 WORDS WERE FOUND  DIRECTLY ON THE TAPE
       443 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
       321 WORDS WERE FOUND  TO BE MISSING
      2253 WORDS WERE FOUND  TO BE DUPLICATES

IN THIS BATCH
      1066 WORDS WERE FOUND  DIRECTLY ON THE TAPE
```

Figure 10. Word-For-Word Lookup Statistics: Cybernetics Before Updating

in Section 12 below. It was felt that translating the text with that high a percentage of missing words would not give a fair picture of the capabilities of our translation program. On the other hand it was also recognized that supplying the dictionary with all the missing words would be equally misleading. As a compromise, therefore, the missing words were supplied only for the first two batches, showing the operation of the program under both sets of conditions. Figure 11 shows the results after the dictionary was updated. The number of missing words in the first two batches was significantly reduced 86%. Those in later batches were reduced only by a small percentage (15% to 20%). Conversely, in the later batches, the number of words found during the lookup directly on the tape increased by 2.2 - 3.8%, the number of words found after stem-ending analysis increased by 4.5-7.4%. For the future, we can envision interesting experiments based on the clustering of words in different types of text.

9. Dictionary Print Program

One of the programs, the output of which is used in the updating of our machine dictionary is the dictionary print. To speed up its operation, the program has been double buffered and converted to STL's System B. In addition, the output format has been changed in the interest of greater efficiency and economy.

The original dictionary listing program listed each of the 108 grammar code bits of the bit pattern code described in Reference 1. The 108 separately printed bits took up an entire line of printout, making the dictionary unwieldy due to its bulk. The print program was therefore rewritten, both to reduce the computer time required for dictionary updating and to reduce size of the dictionary printout.

The bit pattern printout was condensed into an octal pattern corresponding to the three octal computer words it actually takes up in core storage. This was found to be acceptable to the linguists and lexicographers; it also afforded a better survey of the changes in adjacent entries.

This condensation of the dictionary format reduced the cost of printing on peripheral equipment by more than 50%.

```
                                        GLOSSARY LOOKUP
                                  EDITED TEXT CORPUS L  CYBERNET
IN THIS BATCH
        1286 WORDS WERE FOUND  DIRECTLY ON THE TAPE
         409 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
          46 WORDS WERE FOUND  TO BE MISSING
        2351 WORDS WERE FOUND  TO BE DUPLICATES
IN THIS BATCH
        1312 WORDS WERE FOUND  DIRECTLY ON THE TAPE
         387 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
          49 WORDS WERE FOUND  TO BE MISSING
        2344 WORDS WERE FOUND  TO BE DUPLICATES
IN THIS BATCH
         934 WORDS WERE FOUND  DIRECTLY ON THE TAPE
         350 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
         222 WORDS WERE FOUND  TO BE MISSING
        2586 WORDS WERE FOUND  TO BE DUPLICATES
IN THIS BATCH
        1118 WORDS WERE FOUND  DIRECTLY ON THE TAPE
         489 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
         281 WORDS WERE FOUND  TO BE MISSING
        2204 WORDS WERE FOUND  TO BE DUPLICATES
IN THIS BATCH
        1045 WORDS WERE FOUND  DIRECTLY ON THE TAPE
         356 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
         175 WORDS WERE FOUND  TO BE MISSING
        2516 WORDS WERE FOUND  TO BE DUPLICATES
IN THIS BATCH
        1103 WORDS WERE FOUND  DIRECTLY ON THE TAPE
         463 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
         273 WORDS WERE FOUND  TO BE MISSING
        2253 WORDS WERE FOUND  TO BE DUPLICATES
IN THIS BATCH
        1093 WORDS WERE FOUND  DIRECTLY ON THE TAPE
         487 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
         249 WORDS WERE FOUND  TO BE MISSING
        2263 WORDS WERE FOUND  TO BE DUPLICATES
IN THIS BATCH
        1079 WORDS WERE FOUND  DIRECTLY ON THE TAPE
         412 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
         212 WORDS WERE FOUND  TO BE MISSING
```

Figure 11. Word-For-Word Lookup Statistics: Cybernetics After Updating First Two Batches

Figure 12 shows a page of the new dictionary printout, displaying not only the Russian, English, and grammar code, but also the idiom, semantic and stem-ending analysis codes.

10. Grammar Code Sort

The entire RW machine dictionary was sorted by grammar codes in order to evaluate possibilities of space-saving in the future use of grammar codes.

The grammar code format used in RW Division's MT dictionary was chosen for ease of internal computer handling and to ease our lexicographers' task in filling in grammar-coding forms. It is, however, a space-consuming code, occupying 108 bit positions, buying convenience of use at the expense of space. In the space occupied by the grammar code, in theory, 10^{32} different grammatical configurations could be accommodated, clearly much more storage space than is required by existing configurations.

As dictionary space grew more precious, it became desirable to know the number of different grammar codes extant in our glossary, in order to be able to estimate the minimum space required to accommodate the number of grammar codes that can be expected.

A sort of the dictionary conducted when it contained 22,538 entries showed 2098 different grammar codes. These codes could be stored in condensed table-lookup form requiring no more than twelve bit positions each, instead of the present 108.

As the dictionary has mainly been compiled from text, its grammatical composition gives an interesting insight into the grammatical structure of the text, as shown by the following tabulation by parts of speech derived from the sort:

9896	nouns
7255	modifiers (adjectives and participles)
2885	predicates
1370	infinitives
305	adverbs and particles
198	gerunds
30	conjunctions
15	kotoryi forms

There are 205 homographs.

Figure 13 shows a page of the output of the grammar code sort.

NIKTO	NOBODY		400000000000	010200000000	000000000000	J	000	8887		5	00
NIKULINA	NIKULINA		400000000000	000120040000	230000000000	A	000	10605		8	40
NIKFI	NIKFI		400000000000	000200000000	020000000000			3372	1	5	40
NIM	IT+S+HIM+S+THEM+P+		400000000000	010000012020	324000000000		1000	3370		3	00
NIMI	THEM		400000000000	010400000002	334000000000			3369		4	00
NIMFY	NYMPH		400000000000	000010000100	230000000000			10606		5	40
NISKOL+KO	NOT AT ALL		004000000000	000000000000	000000000000	H	000	9704		9	00
NITEVIDNOGO	THREADLIKE+FILIFORM		100000000000	004024040000	000000000000			9705	110	8	01
NITEJ	FIBERS		400000000000	000000000040	010000000000			3360	64	3	77
NITI	WIRE		400000000000	000011000510	010000000000			3360	24	3	77
NITRAT	NITRATE		400000000000	000200200000	020000000000			3362	2	6	47
NITRATA	NITRATE		400000000000	000020000000	020000000000			3362	7	6	47
NITRATE	NITRATE		400000000000	000000002000	020000000000			3362	32	6	47
NITRATNOGO	NITRATE		100000000000	004024000000	000000000000			10607	110	7	01
NITRATREDUKTAZY	NITRATE REDUCTASE		400000000000	000010000110	210000000000			10608	40	14	74
NITRIFIKACII	NITRIFICATION		400000000000	000011000400	210000000000			10609	22	11	74
NITROBENZOL	NITROBENZENE		400000000000	000200200000	020000000000			3361	2	11	47
NITROBENZOLA	NITROBENZENE		400000000000	000020000000	020000000000			3361	7	11	47
NIT(ATYE	FILAMENTOUS		100000000000	004000000110	000000000000			9706	85	6	01
NIT+	FIBER+WIRE		400000000000	000100020000	010000000000			3360	56	3	77
NIX	THEM		400000000000	010400000041	334000000000			3359		3	00
NI(EGO	NOTHING	I	400000000000	000040100000	004000000000	J	A000	8748	42	5	56
NI(EM	NOTHING		400000000000	100000012000	204000000000	J	000	3351		5	00
NI(IPOROVI(EM	NICHIPCROVICH		400000000000	000000010000	260000000000	A	000	10610		13	40
NI(TO$NA	NEGLIGIBLE		200000000000	111000000000	100000000000			3350	9	7	03
NI(TO$NO	NEGLIGIBLE		200000000000	511000000000	000000000000			3350	52	7	03
NI(TO$NOE	NEGLIGIBLE		100000000000	004040100000	000000000000			3350	101	7	01
NI(TO$NYJ	NEGLIGIBLE		100000000000	004200200000	000000000000			3350	94	7	01
NIWETA	POVERTY		400000000000	000100000000	000000000000			6516		6	40
NO	BUT		010000000000	140000000000	000000000000			3349		2	03
NOVATORSKOM	INNOVATORY		100000000000	004000001200	000000000000			13263	65	9	01
NOVA=	NEW		100000000000	004100000000	000000000000			3335	97	3	01

Figure 12. Sample Page of Dictionary Printout

THE GARVIN RUSSIAN GRAMMAR CODES PAGE 57

1	2	3	4	5	6
1680	13828	4	4000000000C0C	C0C0C0000C040	C00CC0CC0CCC
1681	13965	137	400000000000C	C0C0C0000C040	C040C00CC0CC
1682	14018	53	400000000000C	C0C0C0C0C040	C10CC0CCC0CC
1683	14329	311	400000000000	C0000000C040	C20CC0CC0CCC
1684	14542	213	400000000000	C000C0CC0040	210CCCCCCCCC
1685	14543	1	400C00000000	C000C000C040	220CC00C0CCC
1686	14544	1	400000000000	C000C0000040	230CC00CCCCC
1687	14546	2	400000000000	C000C000C040	320CC00C0CCC
1688	14547	1	4C0000000000	C000C000C044	CC0CC00C0CCC
1689	14548	1	400000000000	C000C0000044	004CCCCCCCCC
1690	14551	3	40000000C0C0C	C000C000C044	010C00CC0CCC
1691	14562	11	400000000000	C0C0C0C0C044	02CCC00C0CCC
1692	14563	1	40000000000C	C000C0000C44	110CCCCCCCCC
1693	14565	2	400000000000	C000C0C0C044	210CC00C0CCC
1694	14566	1	400000000000	C000C0000C44	260CCCCCCCCC
1695	14567	1	400C00000000	C000C0C00044	300CC00C0CCC
1696	14574	7	400000000C0C	C0C000000C44	310CC0CCCCCC
1697	14617	43	400000000000	C000C0000C44	320CC00C0CCC
1698	14619	2	400000000000	C000C0C00045	00CCC00C00C0
1699	14620	1	400000000000	C000C0000C50	320CC00C0CC0
1700	14624	4	400C00000000	C000000000100	004CC0CC00CC
1701	14627	3	400000000000	C000C0C0C100	020C00CC0CCC
1702	14629	2	400000000000	C000C0000100	220CC0CC0CCC
1703	14653	24	400C00000000	C000000000100	320CC00C0CC0
1704	14654	1	40000000C0C0C	C000C0000104	020C00C0CCCC
1705	14655	1	400000000000	C000000C0110	C0C000C0CCCC
1706	14664	9	400000000000	C000C0C0C110	C04C00CC0CCC
1707	14667	3	400000000000	C000C0000110	C10CC0000CC0
1708	14908	241	400C00000000	C0C0C0C0C110	020CC0CC0CCC
1709	14909	1	400000000000	C00000000110	034000000CC0

Column	Content
1	Number of the grammar code
2	Cumulative total of dictionary entries
3	Number of entries with the specific grammar code
4-6	Grammar code expressed in octal form

Figure 13. Sample Page of Grammar Code Sort

11. Dictionary Duplication Discriminator

One of the problems associated with the updating of our machine dictionary was the timelag between successive dictionary printings, resulting from the lexicographers' work, keypunching and verifying of the entries, checking, updating and printing of the new dictionary.

Due to this time lag, dictionary entries were sometimes updated, or necessary corrections, more than once, because the printout of a previously updated dictionary was not available by our lexicographers. In order to purge the dictionary from the resulting duplicate entries, a dictionary duplication discriminator program was written, the purpose of which was to sort all dictionary entries and in case of duplicate entries retain only one for the updated dictionary. For checking purposes, the removed duplicate entries are listed separately.

12. Statistics of the Word-For-Word Lookup

The RW dictionary lookup works by successive batches of 4092 words each. In every batch the words are sorted in alphabetic order and compared against the dictionary. The dictionary lookup includes the stem-ending analysis of a text word; it is therefore not necessary to place all of the paradigmatic forms of a word into the dictionary, but only as many forms as are required to insure the success of the stem-ending analysis. The lookup of a given word may thus have any one of four results:

a. the word in its current spelling is found in the dictionary;

b. the grammar code and English equivalent of the word are derived from the dictionary after stem-ending analysis;

c. the word cannot be found in the dictionary even after stem-ending analysis; or

d. the word is identical in spelling with the previous word that was looked up.

The fourth result, identity with a previously found word, was made part of the word-for-word lookup for the purpose of saving computer time. This saving was effected thanks to not having to look up a given word form more than once.

Statistics were kept of the four above-mentioned categories of lookup results. The duplicate category (words identical with previously looked-up words) consistently included between 2300 and 2600 words for each batch of 4092 (56-64%), but our data-gathering scheme did not record to which of the other three categories they belonged.

Figure 14 shows an example of a statistical printout. The small number of missing words in this sample is attributable to the fact that a search had previously been made on the text, and that our lexicographers believed that they had supplied one form from every word paradigm represented in the new-word list. That some missing words remained nevertheless, is due to keypunching errors, misprints and some gaps in our stem-ending procedure.

Incidentally, the fact that three times as many words were found directly on tape as were found by stem-ending analysis is of interest. No single explanation can be suggested.

13. New Syntax Flowcharts

The majority of the flow charts for the syntax program had not been changed since the publication of Reference 1 and were no longer an accurate guide for improving the program or checking out changes that had been made. For some portions of the program no flow charts had been drawn at all and existing flowcharts had not been updated consistently enough to provide an adequate record of the numerous additions and deletions made since they were first drawn up. For the above reasons, and in view of the addition of new staff members unfamiliar with the development of the translation program, it was decided to document the running program in detail in order to insure continuity and to have a more efficient tool for improving and debugging the syntax. A considerable amount of documentation and concurrent checking of program logic has by now been accomplished. It has proved useful in exposing inconsistencies in coding responsible for previously unexplained errors in translation.

```
                    GLOSSARY LOOKUP
                EDITED TEXT CORPUS E 09/02/62

IN THIS BATCH
        1106 WORDS WERE FOUND  DIRECTLY ON THE TAPE
         334 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
          50 WORDS WERE FOUND  TO BE MISSING
        2602 WORDS WERE FOUND  TO BE DUPLICATES

IN THIS BATCH
        1249 WORDS WERE FOUND  DIRECTLY ON THE TAPE
         382 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
          39 WORDS WERE FOUND  TO BE MISSING
        2422 WORDS WERE FOUND  TO BE DUPLICATES

IN THIS BATCH
        1159 WORDS WERE FOUND  DIRECTLY ON THE TAPE
         366 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
          66 WORDS WERE FOUND  TO BE MISSING
        2501 WORDS WERE FOUND  TO BE DUPLICATES

IN THIS BATCH
         153 WORDS WERE FOUND  DIRECTLY ON THE TAPE
          35 WORDS WERE FOUND  AFTER STEM ENDING ANALYSIS
          12 WORDS WERE FOUND  TO BE MISSING
         381 WORDS WERE FOUND  TO BE DUPLICATES
```

Figure 14. Word-For-Word Lookup Statistics: Botany After Updating

14. Changes in the Syntax Program

The nominal-blocking and predicate-blocking routines have been revised because the operation of the Translation Error Detector requires the comparison of the first words of a predicate block or the first word of a nominal block of the machine translation with a corresponding word in the human translation. The original syntax program did not recognize the boundary between adjacent syntactically identical blocks, and hence provided no means of finding the first word of the second of the two blocks for purposes of matching with the human translation. This could lead to serious errors in matching. Therefore, a block-starting code was devised which identifies the first word of a block even if another block of the same syntactic structure precedes, thus enabling the program to separate two adjacent syntactically identical blocks.

15. Fail-Safe Devices

Since the current contract called for translation of larger amounts of text than we were previously set up to process by our translation program, the question of fail-safing acquired greatly increased significance. A number of appropriate modifications were therefore introduced into the program.

The first of these was designed to detect long loops through the syntax program. This was done by inserting a counter into a grammar-code checking subroutine which is used very frequently by all sections of the syntax program. When the counter registers an abnormally large number of entrances into the subroutine, the program assumes that it has gone into a long loop. Processing of the problem sentence is halted, and a record is made of that sentence and the program locations which have been altered during the processing of the text.

Other modifications were required for dealing with certain types of overflow conditions which had not previously been encountered. In the original syntax program, an arbitrary limit of 100 Russian words had been set as the maximum sentence length which could be handled by the

program; the two sentence storage locations (bins) were accordingly limited to 1900 computer words each (19 computer words = 1 Russian word). The processing of a greater diversity of texts under the current contract called for a subroutine capable of handling a sentence of any length. A modification was therefore incorporated to recognize sentences of over 100 words and break them into smaller segments at major punctuation marks, which allowed the continued use of existing sentence bins, rather than expanding to another arbitrary figure.

In addition to providing for overflow from the sentence bins, we also made provisions for overflow of two smaller bins.

One of these is the bin used for storing the addresses of word blocks waiting for rearrangement while the rearrangement parameters are being computed. This bin had previously been limited to 20 words, which proved inadequate for rearrangement of very large subject and object blocks. It was expanded to 30 words and an overflow check was installed to prevent rearrangement of blocks exceeding the new limit.

An overflow check was also provided for the skip bin. This is a string of cells used to store syntactically inert words that have been removed from the sentence before the major syntax passes in order not to interfere with the searches. When an overflow is found, the sorting of skip words is halted, a signal is stored to indicate overflow, and the processing of the sentence goes into the next pass.

In addition to these modifications, an aid to checking out program improvements was provided in the form of a list of all the memory locations that have been changed during processing of a text. Before beginning to process the first sentence of a text, the program is stored on a tape. When a problem sentence (one in which an overflow or loop has been detected) is encountered, the program as it stands in core is compared with the program tape and all memory locations which differ from those originally recorded on the tape are written onto another tape containing all executive output.

16. Translation

All or part of the keypunched text in the following fields have been machine-translated:

soil science
Pavlovian psychology
biology
botany
education
fiction
cybernetics

The Translation Error Detector routine has been applied to text in the fields of biology and botany.

A sample page of the "vertical" translation output is shown on Figure 15.

17. Concordance

A routine has been written and is currently being checked out to produce a concordance of those words of the machine translation which the dictionary transformation routine has successfully matched with the human translation. Each line will have the same format as that shown previously in the record of word matches (called the "unsorted concordance" in Reference 2), but with the following difference: instead of showing the several English equivalents stemming from the dictionary lookup, we show only the single matching translation. This routine will be of major assistance in the conduct of our semantic studies.

A sample page of the output is shown on Figure 16.

```
CORPUS L 10/29/62        E Q            225

B 710          200000 1  5  V                          IN                                          I        644
.......1. . . ........................ . ...... ............................ . ...... ............1111...111...11.1

B 711 30 74    201000 2  6  DRUGIX                     OTHER+D+                                       1     4862
..X...... . . ........................ . ...... ...........................X . ...... ...........................

B 8 1          201000 0  7  SLU(A=X          .         CASES+OCCASIONS+EVENTS+INCIDENTS                     1451
X........ . . ........................ . ...... ...........................X . ...X.. ...........................

B 8 2 40 41    020000 0  8  OBRAZOVYVALIS*      WERE          PRODUCED     + FORM+EDUCATE          *        3240
.1....... . . ........................ . .11... ............................ . .....11 ............11.1..........1...

B 8 3          001002 2  9  NEBOL*)IE                  SMALL                                                3565
..X...... . . ........................ . .....X ......................X..X... . ...... ...........................

B 8 4          001002 0 10  VYROSTY                    PROTUBERANCES      +OUTGROWTHS              *        0190
1........ . . ........................ . ...... .....................1..1... . ...1.. ...........................

B 8 5          200000 1 11  V                          IN                                          I        644
.......1. . . ........................ . ...... ............................ . ...... ............1111...111...11.1

B 8 6          201000 2 12  VIDE                   THE FORM                                                  609
X........ . . ........................ . ...... ..................X......... . ...X.. ...........................

B 8 7          201200 2 13  KONUSOV                 OF CONES                                       *        9644
1........ . . ........................ . ...... .......................1.... . ...1.. ...........................

B 8 8          000000 0  4  ILI                        OR                                          H I      5181
.....1... . . ........................ . ...... ............................ . ...... ...........................

B 8 9          001000 2  1  VYROSTY                    PROTUBERANCES      +OUTGROWTHS              *        0190
1........ . . ........................ . ...... .....................1..1... . ...1.. ...........................

B 9 1 77       001200 2  2  BOL*)EGO           OF THE LARGER                                          CA       4606
..X...... . . ........................ . .....X .....X...................... . ...... ...........................

B 9 2          001200 0  3  RAZMERA                    DIMENSION                                                1767
```

Figure 15. Sample Page of "Vertical" Translation Output

```
587RS 590ES  12MA  18TO  66QU  OSW   ODF   OMISM
VB22 1  3 420000 0  ODNOVREMENNO                  SIMULTANEOUSLY
VB22 2  4 420000 0  VOZNIKAET          .         ARISES + RISES
VB22 3  5 401000 2  MNO$ESTVO                 THE MANIFOLD
VB22 4  6 401200 2  MIKROKLETOK       ,         OF MIKROKLETOK
VB23 5  2 301200 0  RAZMEROV                     DIMENSIONS
   SAME              ..   3
   TIME              ..   4
   THERE             ..   5
   APPEARED          ..   6
   NUMEROUS          ..   7
   MICROCELLS        ..   8
   SIZE              ..  19
   FIG               ..  20

588RS 591ES   6MA  11TO  54QU   OSW   ODF   OMISM
VB23 7  2 000000 0  KAK                          AS+HOW                                  I B
VB24 3  6 001000 2  PUTEM                     BY MEANS                                   H A
VB24 4  7 001200 2  OTPO(KOVANI=           OF THE BUDDING-OFF
VB24 5  8 001200 2  U(ASTKOV                  OF PARTS
   IT                ..   1
   THAT              ..   4
   BUDDING           ..  13
   OFF               ..  14
   PORTIONS          ..  16
   FIG               ..  24

589RS 592ES   7MA  11TO  63QU   OSW   ODF   OMISM
VB25 2  2 400000 0  POSLEDNIX                    LAST+D+                                 *H
VB25 4  4 601000 2  RANNIX                       EARLIER
VB26 2  9 420000 0  EDINI(NY          ,      ARE SINGLE          +UNIT                   *
VB26 4 11 020000 0  PO=VL=QTS=                   APPEAR
   RARE              ..   3
   AMONG             ..   4
   LATTER            ..   6
   EARLY             ..   9
   BUT               ..  14
   OCCURRED          ..  16

590RS 593ES   6MA  10TO  60QU   OSW   ODF   OMISM
VB26 8  3 001000 0  U(ASTKOV                     PARTS
```

Figure 16. Sample Page of Output of Concordance Routine

18. Detailed Description of the Translation Error Detector

As was stated in Part I, a statistical record is kept by the matching program of the correspondence of the insertion of a number of function words in the machine translation and the human translation. The function words of interest are the three written article forms, 11 prepositions, and 25 verbal auxiliaries. They are listed below:

A. Articles

THE

A, AN

B. Prepositions

OF

IN, INTO

TO

AT

FOR

AFTER

BY

FROM

WITH

ON

C. Auxiliaries

BE, AM, ARE, IS, WAS, WERE, BEEN

HAVE, HAS, HAD

DO, DOES, DID, DONE

SHALL, SHOULD

MUST

MAY, MIGHT

OUGHT

CAN, CANNOT, COULD

WILL, WOULD

For each of these words, we reserve four counters in the computer. Ideally, they will tabulate:

(1) the same insertion has been made in the machine translation as in the human translation (called "corresponding insertion");

(2) an insertion has been made in both the machine translation and the human translation, but the inserted words are not the same (called "non-corresponding insertion");

(3) no insertion has been made in either the machine translation or the human translation (called "corresponding non-insertion");

(4) An insertion has been made in the machine translation where none has been made in the human translation, or conversely, no insertion has been made in the machine translation where one has been made in the human translation (non-corresponding non-insertion).

The four counters are arranged in a matrix as follows:

	Corresponding	Non-Corresponding
Insertion (translation)	1	2
Non-insertion (non-translation)	3	4

In present practice, counters 1 and 2 are as stated above; counter 3 has been established only for the article THE; counter 4 is used for the cases where the human translation uses a word that does not appear in the machine translation, or conversely.

The counters, as was stated above, operate in conjunction with the matching program. When both the machine translation and the human translation of a particular sentence have been successfully matched and brought into core together, the program first searches the human translation for words of interest, and then inspects the machine translation to look for corresponding words. When this search is completed, the program proceeds to the converse: searching the machine translation first and then proceeding to the human translation to look for corresponding words.

We will use the sample sentence shown in Figures 1 and 2 to illustrate this process in more detail. The detailed flowcharts are shown in Appendix C.

From Human Translation to
Machine Translation

This portion of the program starts out with the function words for which the counters have been established, namely, articles, prepositions, and auxiliaries, in that order.

In our sample sentence, the program looks for and finds the first article in the human translation: THE, preceding the word SHAPE. The program then asks if the word following this article was matched with a word in the machine translation. The answer is "no", since the machine translation of the corresponding Russian word is THE FORM. This article is therefore not suited for tabulation by the counter and the program looks for the next occurrence of an article. No further articles are present in this sentence and therefore the program proceeds to the next group of function words.

These are the prepositions. The first preposition is the human translation IN. Again the program reads the next following word (OTHER) and asks whether it is matched with a word in the machine translation. In this case, the answer is "YES". Since the program is now dealing with prepositions it next asks whether the word matched in the machine translation forms part of a prepositional block. To this the answer is likewise "YES" (in the original Russian: V DRUGIX SLU(A=X '. The program now compares the machine translation of the first word of this prepositional block to the preposition found in the human translation, and records their identity (IN = IN). This is therefore a case of corresponding insertion/translation, and counter 1 is increased by one.

The program now reads the next preposition in the human translation, which is another instance of "IN", namely that before THE SHAPE. This portion of the program skips over articles to read the word following the preposition, which then is SHAPE. This word finds no match in the machine translation, and the current preposition is rejected for tabulation just as was the article in the earlier instance.

The next preposition in the human translation is "OF". The following word CONES, has a match in the machine translation, and the corresponding

Russian word is in a Russian genitive nominal block. As in the previous case, we take the first word of this block and compare it to our preposition. They are identical. The insertion of this preposition by the machine translation program has been correct; therefore, counter 1 for this preposition is increased by one.

There is one more preposition in our sample sentence, it is again "OF", and a corresponding insertion is obtained exactly as in the preceding case.

There are no auxiliaries present in the human translation, the human-to-machine portion of the matching program has nothing more to compare.

The second portion of the program therefore goes into effect.

From Machine Translation to Human Translation

This portion of the program looks in turn at all those words of the machine translation that stem from the machine dictionary rather than from insertions, and that have found a match in the human translation. It checks the syntax record of each of these words and retains it only if it is either a verb or the first word of a nominal block. It then checks the insertion record. The first such word is SMALL. This is identified as the first word of a nominal block; no articles or prepositions were inserted before it by the translation program. None appear before the corresponding word of the human translation. We therefore increase by one counter 3 for the article THE (only word for which this counter has been established).

The next matching words OUTGROWTHS and IN are neither verbs nor first words of nominal blocks and are therefore ignored.

The word CONES is not only the first word of a nominal block, but also has OF inserted before it. This insertion, however, has already been accounted for by the human-to-machine portion of the program and is therefore no longer taken into consideration.

OR is neither a verb nor the first word of a nominal block.

PROTUBERANCES meets the same conditions as SMALL before, hence, counter 3 for THE is increased to two.

O	THE	113	44	97	37	A		15	0	0	OF	62	16	0	10
O	INTO	2	1	0	0	IN	19	9	0	1	AT		1	0	0
O	TO	4	7	0	2	FOR	1	2	0	2	AFTER	4	1	0	0
O	BY	1	4	0	2	FROM	2	1	0	0	WITH	6	2	0	0
O	ON	1	2	0	0	IS	4	4	0	6	ARE	1	1	0	0
O	WAS		1	0	0	HAVE				2	MAY		1	0	0
O	WERE		2	0	0	WILL	1	0	0	0					
	PA 23INCREASE														
	PB 2ESTABLISHING														
P	THE	29	41	108	20	A		9	0	3	OF	23	20	0	7
P	IN	16	10	0	2	AT		2	0	0	TO	1	3	0	1
P	FOR	3	1	0	0	AFTER	1	0	0	0	BY	6	15	0	3
P	FROM	3	0	0	0	WITH	8	1	0	0	ON	5	1	0	0
P	IS	1	0	0	0	ARE		1	0	3	WAS	2	2	0	0
P	DID		1	0	0	WERE	2	3	0	1					
	QA 39TOLUIDINE														
	QB 28LEFT														
	QD 22YOUNG														
Q	THE	100	125	227	70	A		14	0	2	OF	99	64	0	6
Q	IN	20	18	0	7	AT		14	0	0	TO	2	0	0	1
Q	FOR	1	1	0	1	AFTER	22	0	0	1	BY	3	8	0	2
Q	FROM	3	1	0	0	WITH	5	25	0	0	ON	7	0	0	0
Q	IS	1	2	0	0	ARE	1	2	0	1	WAS	4	7	0	2

Figure 17. Sample Page of Statistical Printout of Translation Error Detector

LARGER is the first word of a nominal block. OF inserted before it has already been accounted for by the human-to-machine portion, THE has not. The latter word has no match in the human translation and is therefore counted as a non-corresponding non-insertion, which increases counter 4 for THE by one.

DIMENSIONS is skipped because it is neither a verb nor the first word of a nominal block.

FORMED is a verb. The auxiliary WERE has been inserted before it, and no record of it was left by the human-to-machine portion of the program. Counter 4 for WERE is therefore increased by one.

OTHER meets the same conditions as SMALL, hence, counter 3 for THE is increased by still one more, to three.

CASES is skipped because it is neither a verb nor the first word of a nominal block.

This completes the processing of our sample sentence.

We have used the Translation Error Detector to obtain preliminary statistics for text from two of the chosen fields. A sample page of our statistical printout is shown on Figure 17. Figure 18 shows the percentage of correct insertions (translations) for six function words over the entire text processed so far.

Function Word	Total Number of Occurrences	Number of Correct Insertions (Translations)	$\frac{\text{Column 2}}{\text{Column 1}} \times 100$
THE	5704	3625	63.6%
OF	1998	1114	55.6
IN	680	309	45.4
BY	200	41	20.5
TO	147	44	29.9
WAS	128	45	35.2

Figure 18.

19. The "Bôche-Fèchre" (Botch-Fetcher)

The purpose of this program is to provide linguists with a record of the results of the errors (or successes of certain insertion or translation rules. This is done by providing each sentence processed by the Translation Error Detector with one or several error codes. These codes show exactly the increases caused by the sentence in question in each of the four counters provided for each of the 39 function words. Each sentence is thus tagged with a code relating it to one of the 39x4 counters. It now becomes possible to retrieve automatically and print out in their entirety all the sentences in which one of the 39 function words was treated in a particular way, for instance, all the sentences in which OF was incorrectly inserted, as shown by increases in counters 2 and 4. The printout will give not only the sentences, but for each, also information about grammatical packaging, syntactical decisions, idiomatic use, homograph resolution, etc.

20. Chinese-English MT

A survey of problem areas in Chinese-English machine translation was conducted with a view towards the application of the fulcrum approach and of the computer-aided research procedures developed under the present contract to this new field. This survey resulted in a number of tentative conclusions which are discussed in detail in Reference 3.

REFERENCES

(1) "Machine Translation Studies of Semantic Techniques," Technical Report No. 1, C72-1U7, 22 February 1961, Thompson Ramo Wooldridge Inc., Ramo-Wooldridge Div. (Unclassified)

(2) "A Manual for Automatic Dictionary Revision," C79-0U7, 18 August 1960, Thompson Ramo Wooldridge Inc., Ramo-Wooldridge Div. (Unclassified)

(3) "Survey of Problem Areas in Chinese-English Machine Translation," Technical Note No. 1, under Contract NSF-C233 with NSF, "Computer-Aided Research in Machine Translation," 21 August 1962, Thompson Ramo Wooldridge Inc., RW Division. (Unclassified)

APPENDIX A

TRANSLITERATION TABLE

APPENDIX A

TRANSLITERATION TABLE

Cyrillic	Transliteration Linguistic	Transliteration Keypunching	Cyrillic	Transliteration Linguistic	Transliteration Keypunching
А,а	A	A	Р,р	R	P
Б,б	B	Q	С,с	S	C
В,в	V	B	Т,т	T	T
Г,г	G	L	У,у	U	Y
Д,д	D	V	Ф,ф	F	$
Е,е	E	E	Х,х	X	X
Ж,ж	$	S	Ц,ц	C	U
З,з	Z	Z	Ч,ч	(	%
И,и	I	N	Ш,ш	)	W
Й,й	J	@	Щ,щ	W	&
К,к	K	K	Ъ,ъ	/	D
Л,л	L	J	Ы,ы	Y	I
М,м	M	M	Ь,ь	*	F
Н,н	N	H	Э,э	9	G
О,о	O	O	Ю,ю	Q	¤
П,п	P	/	Я,я	=	R

APPENDIX B

KEYPUNCHING INSTRUCTIONS FOR ENGLISH TEXT

APPENDIX B

KEYPUNCHING INSTRUCTIONS FOR ENGLISH TEXT

Table I - Transliteration Code

SYMBOL	CHARACTER(S) TO BE TYPED
. (period)	.
! (exclamation point)	.
? (question mark)	$Q
; (semi-colon)	$,
: (colon)	$.
, (comma)	,
- (hyphen)	—(dash or @ sign, not X-punch)
— (dash)	—(dash or @ sign, not X-punch)
" (quotation marks)	$/
((open paren)	(no space afterwards
) (closed paren)	) no space before
' (apostrophe)	$- $ followed by X-punch
= ≠ > ≥ < ≤ → (equal and unequal signs)	$$
- (minus sign)	#- # followed by X-punch no space afterwards
° (degree)	#.
$ (dollar sign)	$
/ (slash)	/ no space before or afterwards
. (decimal point)	.
... (excerpt)	##
α (alpha)	#A
β (beta)	#B
φ (phi)	#F

KEYPUNCHING INSTRUCTIONS FOR ENGLISH TEXT

Table I - Transliteration Code

(cont'd)

SYMBOL	
γ (gamma)	#G
ψ (psi)	#I
λ (lambda)	#L
μ (mu)	#M
π (pi)	#P
% (percent sign)	PERCENT
+ $\pm$ (plus, plus-minus)	+
An equation that cannot be keypunched	#E
Mixed symbols or numbers that cannot be keypunched	#S

***** Means a new article. Immediately following is the article letter followed by two spaces. Immediately following the two spaces is the article title. If more than one card is needed, ** begins the second and following cards.

****** The authors' names follow the two spaces after the 6 asterisks. If more than one card is needed, ** will be on each of them.

**** Means a new column of text information. This will stand alone on the card and is the indication to the computer that the cards following will form a new column of text.

* Next character was capitalized in the text.

A new line of text always starts a new card; however, when a line of text cannot be completed on a card, ** at the beginning of a card indicates that this card is a continuation of an old line (from the point of view of the printed text). If no space follows the double asterisk, it further means that the word on the previous card is not completed and is being continued.

Three conventions are necessary for headings, paragraphs, and sentences:

1. A heading and subheading will always be indented by four spaces.
2. A paragraph will always be indented by three spaces.
3. A sentence (unless it is the start of a paragraph or heading) will always be preceded by two spaces.

The sentence convention (3) is needed to allow the edit program to decide when a period is the end of a sentence and when it marks an abbreviation.

(A) All cards have a sequence number in columns one through four. These sequence numbers will be ascending within a given document, but not necessarily consecutive. The text information is keypunched from column five on.

(B) Special care should be exercised to have two spaces appear only before the beginning of a sentence, not after every period.

Example: *THE *U.*S.*A.␣*EXPORTS. ␣ is the symbol for one space.

(C) Be very careful to distinguish the letter O (here written as ∅) from the number 0, and the letter I from the number 1.

(D) The only way to separate one word from another is to allow a space or to keypunch a hyphen - or a dash —
Since open parenthesis (plus-sign + and minus-sign - do present problems, the rule is to attach them to the following word, i.e., leave a space before but do not space after (+-
The closing parenthesis) similarly is attached to the previous word, so do not leave a space before)

(E) - (hyphen, not minus-sign), and — (dash) are keypunched as an @ sign, not an X-punch, and may have spaces around them.

(F) Punch all dimensions as one word: g/cm^3 meqH $^{\circ}C$

(G) Leave out all superscripts or subscripts: Al_2O_3 is *AL*∅

(H) Do not punch footnotes or bibliographies, but do punch (Table 1) (see Tabl. 5B) (Fig. 11A) (1,pp.180-250)

(I) Do not punch tables, their headings, or their labels.

(J) At the end of an article, do not punch:

Received Nov. 25, 1958

EXAMPLES

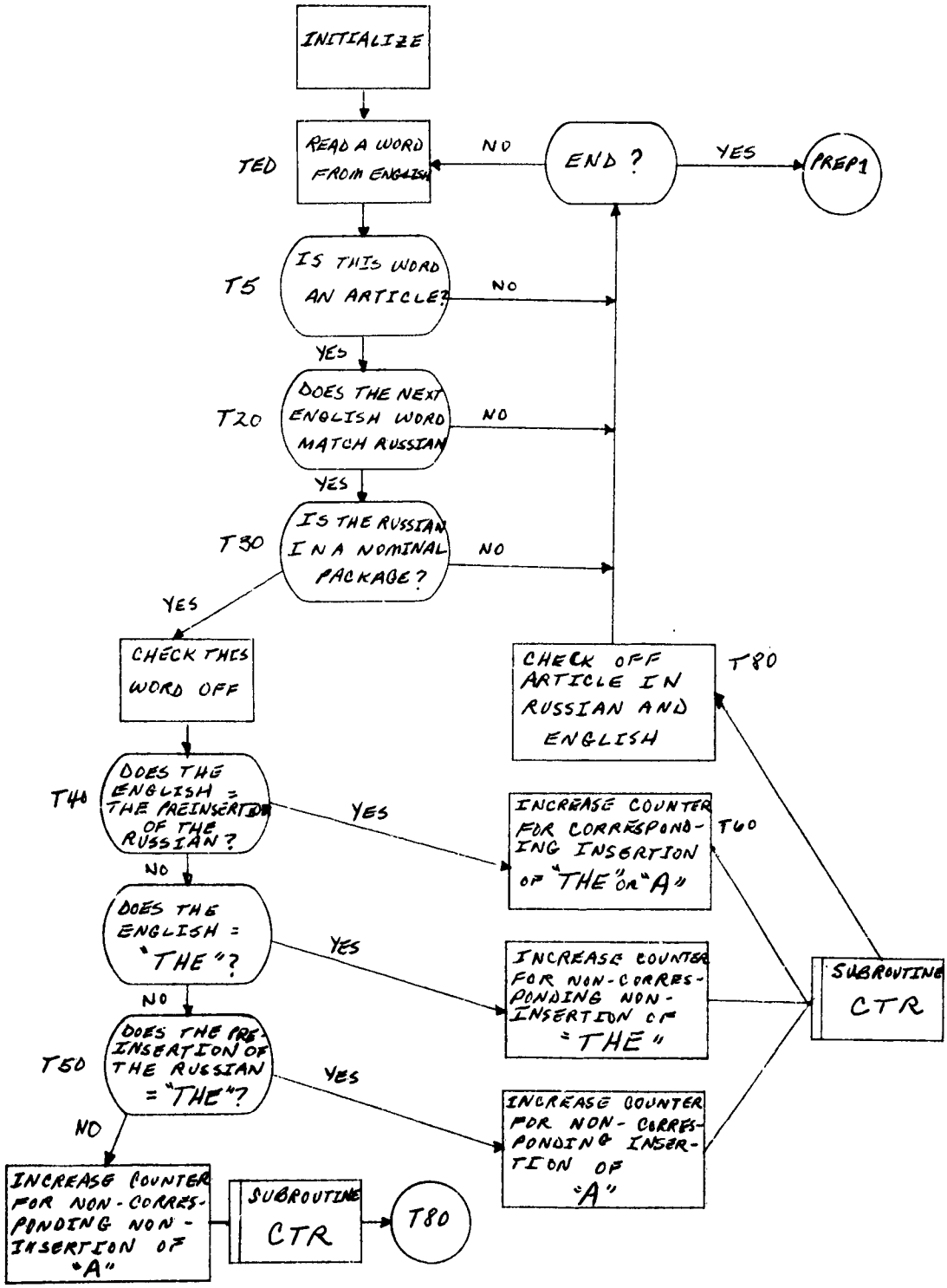
1.	A.M.	will be keypunched as	*A.*M.
2.	Ca^{++}(24.31 meq H)		*CA△(24.31△MEQ*H)△
3.	79%- 84%(Table 1).		79△PERCENT—△84△PERCENT△(*TABLE△1).
4.	(1·2 g/cm^3);		△(1.2△G/CM)$,
5.	Mg^{++}(2.14),Na^{+}(0.73)		*MG△(2.14),△*NA△(0.73)△
6.	±0·8-10°C		+0.8—10△#.*C
7.	30- 40 cntr/ha in 1952-53		30—△40△CNTR/HA△IN△1952—53
8.	339, 333 and 340 g/m^2.		339,△333△AND△340△G/M.
9.	Iron-humus (Fig. 1).		*IRØN-HUMUS△(*FIG.△1).△△
10.	non-cultivated		NØN—CULTIVATED
11.	below -25.0°C		BELØW△#-25.0△#.*C
12.	(1,pp. 254-256)		(1,△PP.△254—△256)△
13.	"John's,— and Williams',"		$/*JØHN$-S,—AND△*WILLIAMS$.-,$/

P.S. In all these examples, except for No. 11, all hyphens - or dashes — are keypunched as an @ sign. No. 13 has both kinds.

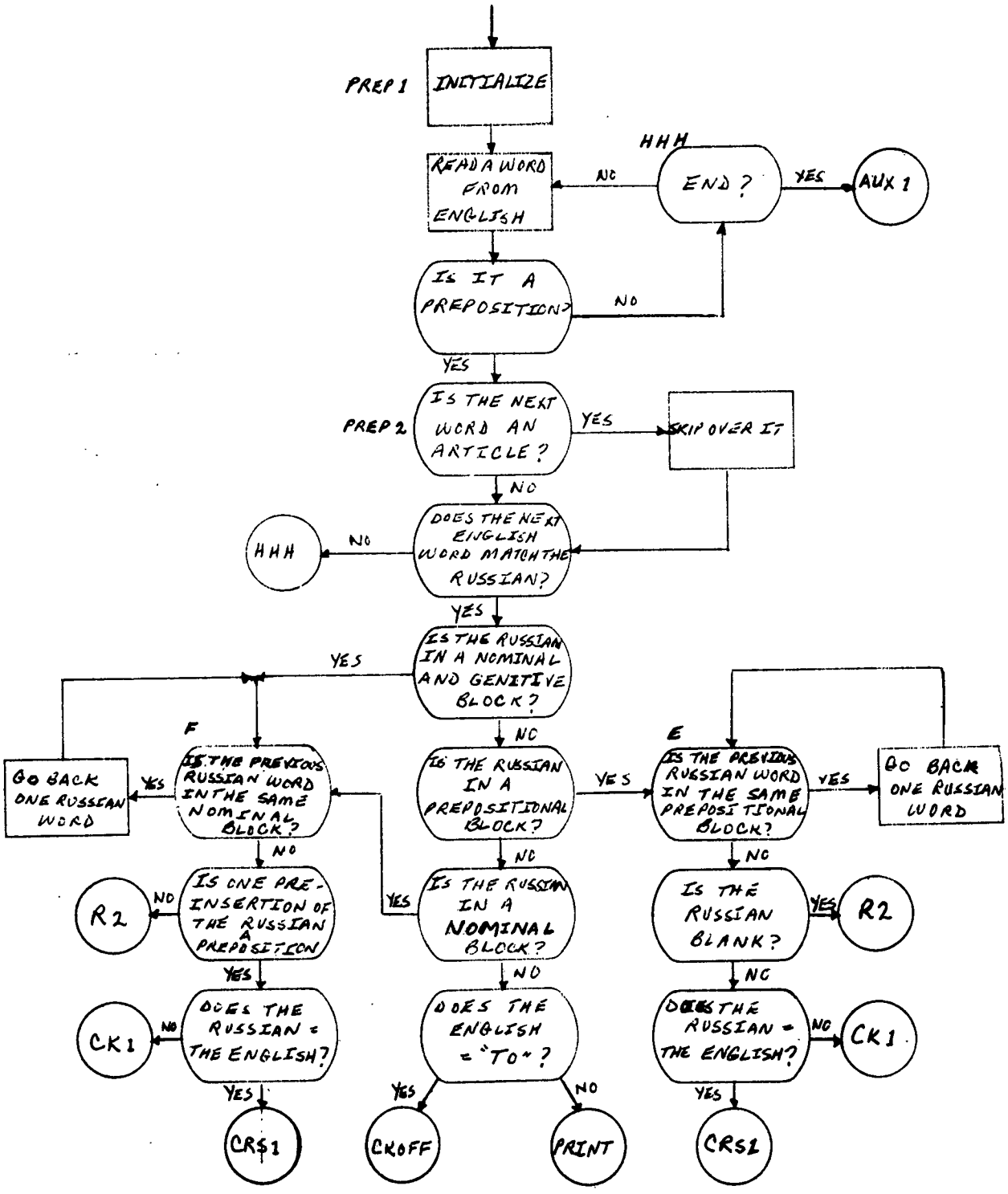
4-15-62
Gerhard Reitz

APPENDIX C
FLOWCHARTS OF
TRANSLATION ERROR DETECTOR

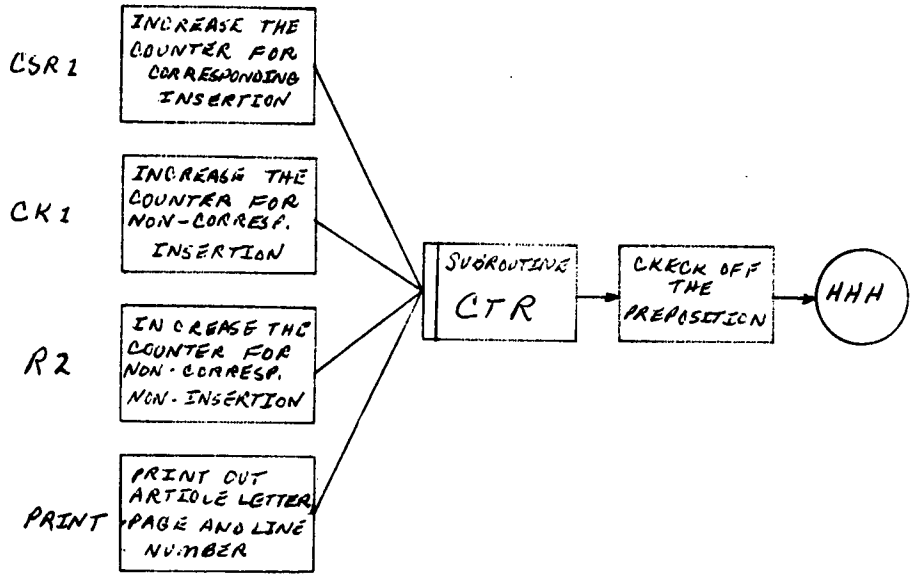
Translation Error Detector

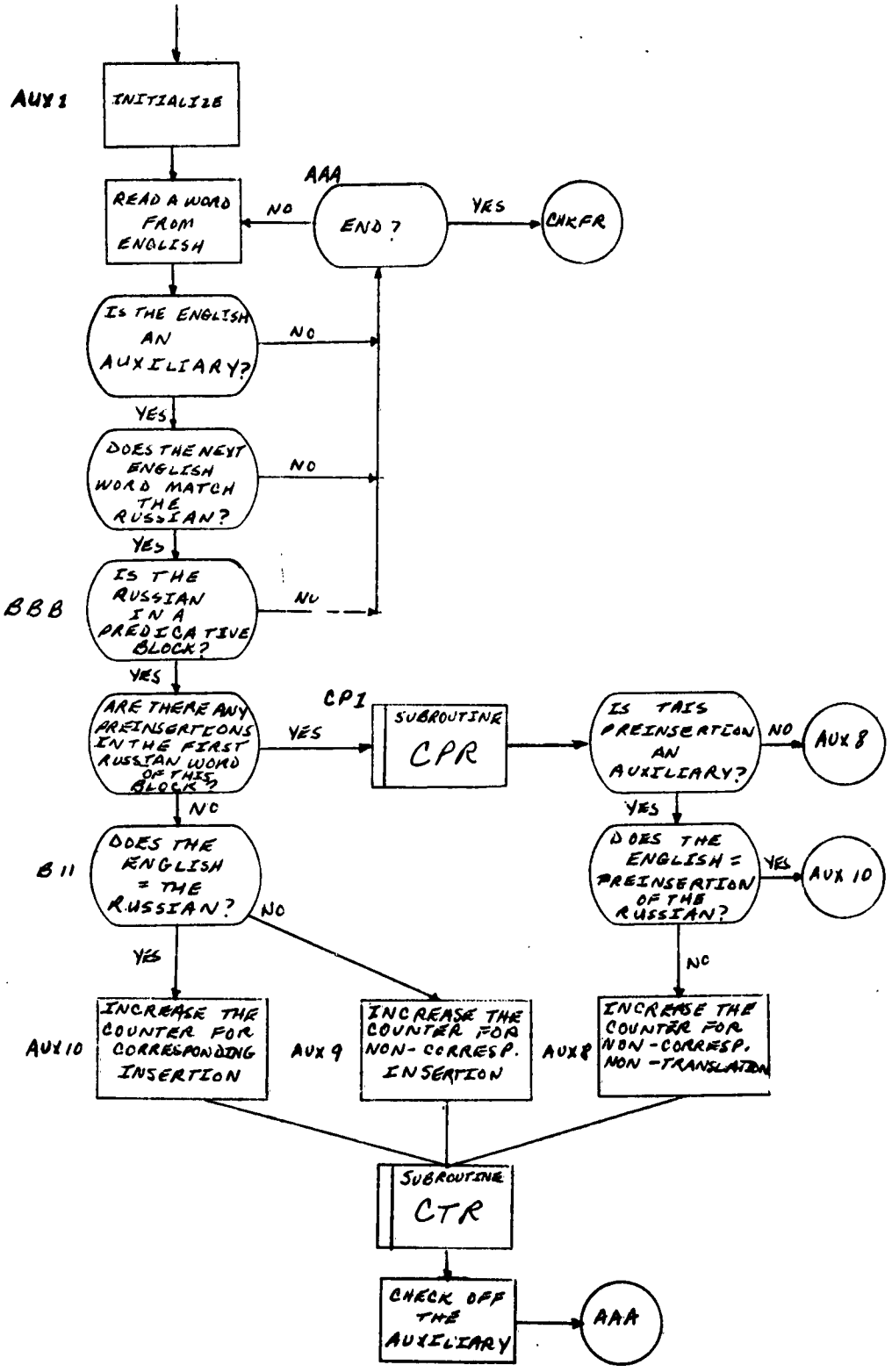


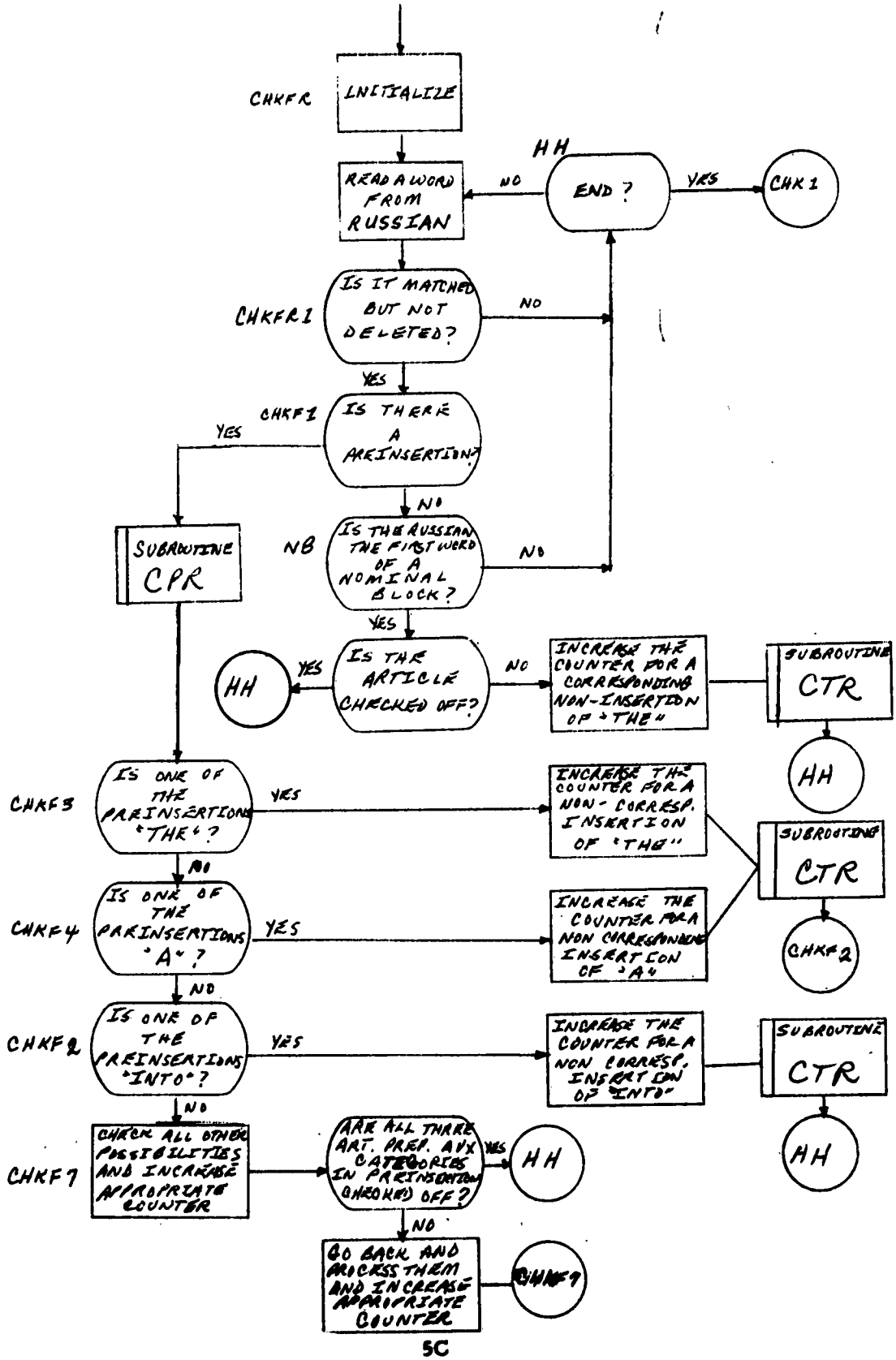
Translation Error Detector

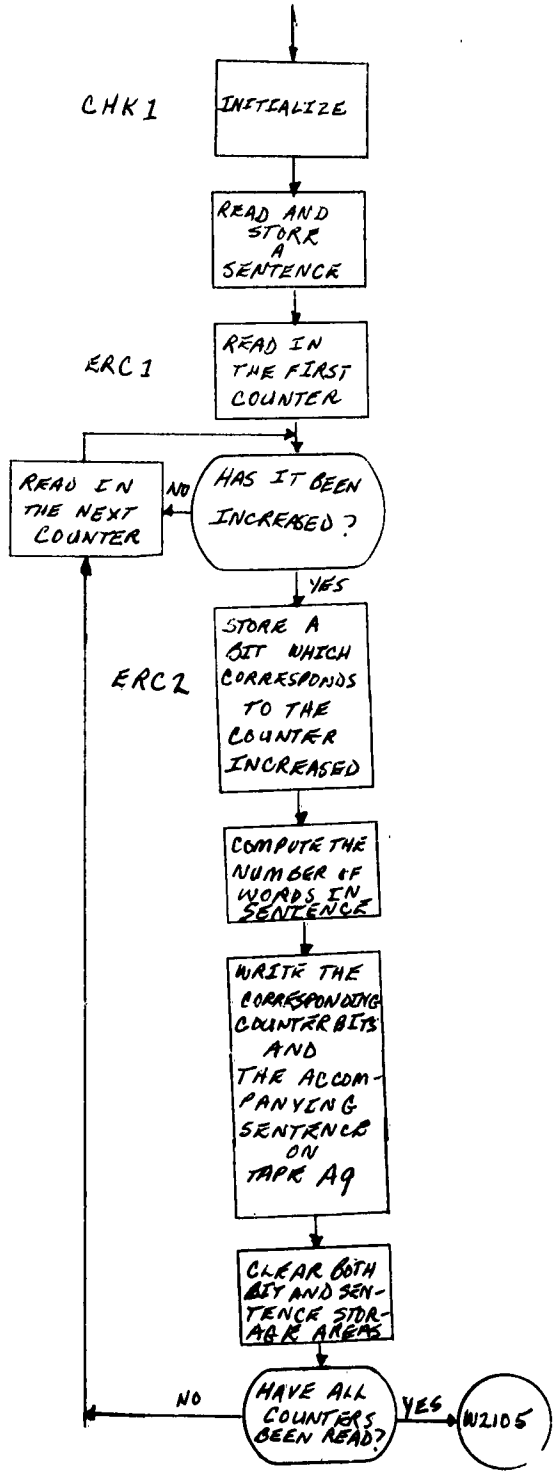


Translation Error Detector



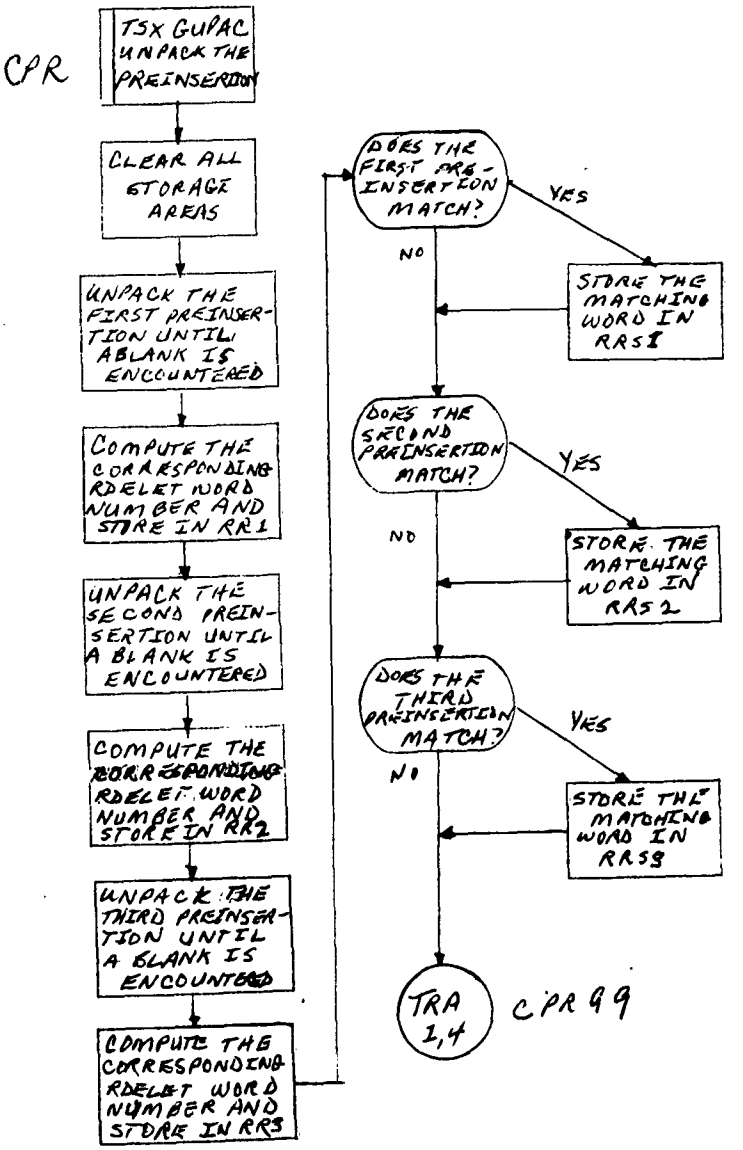

AUX 1
INITIALIZE
AAA
READ A WORD FROM ENGLISH
NO
END ?
YES
CHKFR
IS THE ENGLISH AN AUXILIARY?
NO
YES
DOES THE NEXT ENGLISH WORD MATCH THE RUSSIAN?
NO
YES
BBB
IS THE RUSSIAN IN A PREDICATIVE BLOCK?
NO
YES
CP1
ARE THERE ANY PREINSERTIONS IN THE FIRST RUSSIAN WORD OF THIS BLOCK?
YES
SUBROUTINE CPR
IS THIS PREINSERTION AN AUXILIARY?
NO
AUX 8
YES
NO
B 11
DOES THE ENGLISH = THE RUSSIAN?
NO
YES
DOES THE ENGLISH = PREINSERTION OF THE RUSSIAN?
YES
AUX 10
NO
AUX 10
INCREASE THE COUNTER FOR CORRESPONDING INSERTION
AUX 9
INCREASE THE COUNTER FOR NON-CORRESP. INSERTION
AUX 8
INCREASE THE COUNTER FOR NON-CORRESP. NON-TRANSLATION
SUBROUTINE CTR
CHECK OFF THE AUXILIARY
AAA

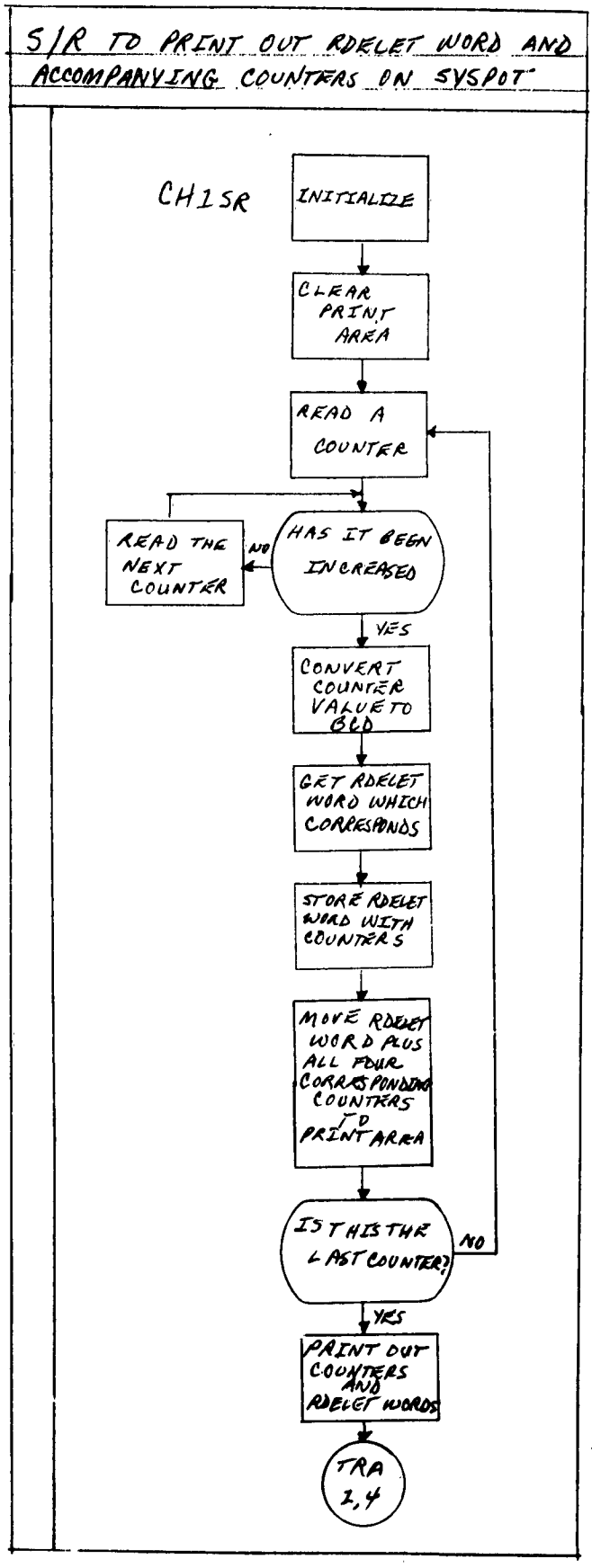

CHKFR
INITIALIZE
HH
READ A WORD FROM RUSSIAN
NO
END ?
YES
CHK 1
CHKFR1
IS IT MATCHED BUT NOT DELETED?
NO
YES
CHKF1
IS THERE A PREINSERTION?
YES
NO
SUBROUTINE CPR
NB
IS THE RUSSIAN THE FIRST WORD OF A NOMINAL BLOCK?
NO
YES
HH
YES
IS THE ARTICLE CHECKED OFF?
NO
INCREASE THE COUNTER FOR A CORRESPONDING NON-INSERTION OF "THE"
SUBROUTINE CTR
HH
CHKF3
IS ONE OF THE PREINSERTIONS "THE"?
YES
INCREASE THE COUNTER FOR A NON-CORRESP. INSERTION OF "THE"
NO
CHKF4
IS ONE OF THE PREINSERTIONS "A"?
YES
INCREASE THE COUNTER FOR A NON CORRESPONDING INSERTION OF "A"
SUBROUTINE CTR
CHKF2
NO
CHKF2
IS ONE OF THE PREINSERTIONS "INTO"?
YES
INCREASE THE COUNTER FOR A NON CORRESP. INSERTION OF "INTO"
SUBROUTINE CTR
HH
NO
CHKF7
CHECK ALL OTHER POSSIBILITIES AND INCREASE APPROPRIATE COUNTER
ARE ALL THREE ART. PREP. AVX CATEGORIES IN PREINSERTION CHECKED OFF?
YES
HH
NO
GO BACK AND REPROCESS THEM AND INCREASE APPROPRIATE COUNTER
CHKF7

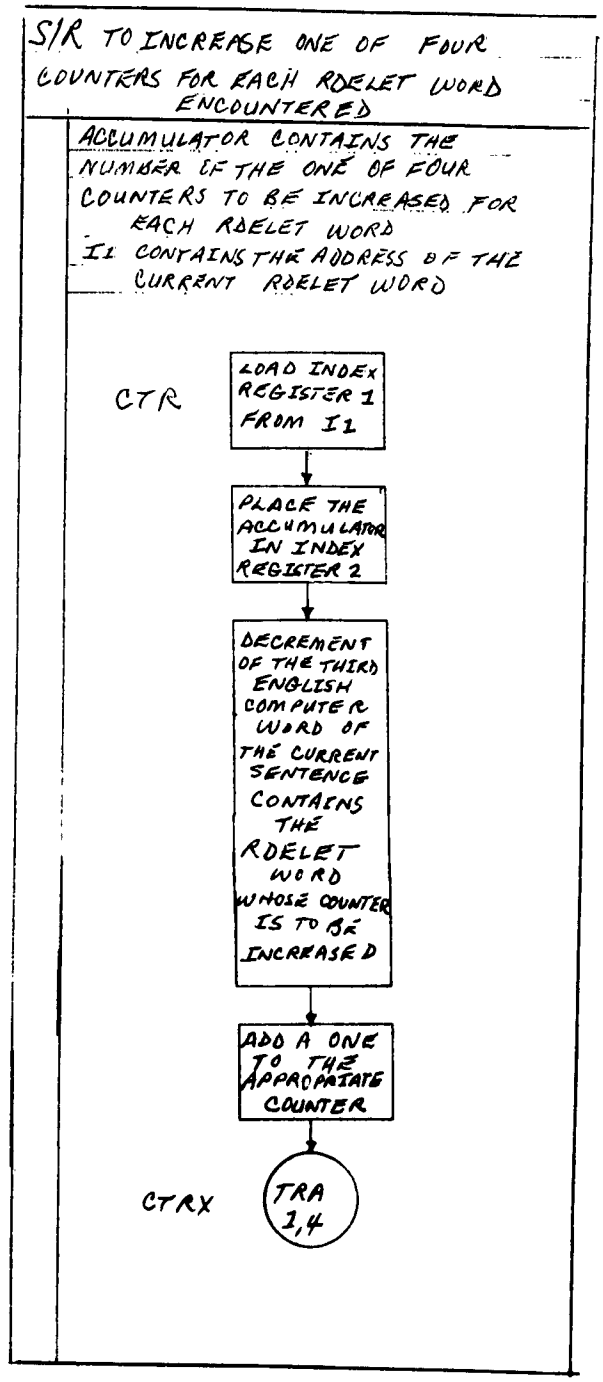

CHK 1
INITIALIZE
READ AND STORE A SENTENCE
ERC 1
READ IN THE FIRST COUNTER
READ IN THE NEXT COUNTER
NO
HAS IT BEEN INCREASED?
YES
ERC 2
STORE A BIT WHICH CORRESPONDS TO THE COUNTER INCREASED
COMPUTE THE NUMBER OF WORDS IN SENTENCE
WRITE THE CORRESPONDING COUNTER BITS AND THE ACCOMPANYING SENTENCE ON TAPE A9
CLEAR BOTH BIT AND SENTENCE STORAGE AREAS
NO
HAVE ALL COUNTERS BEEN READ?
YES
W2105

S/R TO CHECK THE RUSSIAN PREINSERTION FOR A MATCH WITH THE ARTICLE, PREPOSITION AND/OR AUXILIARY

PREINSERTION CONTAINS UP TO THREE BCD WORDS




S/R TO PRINT OUT RDELET WORD AND ACCOMPANYING COUNTERS ON SYSPOT
CH1SR
INITIALIZE
CLEAR PRINT AREA
READ A COUNTER
HAS IT BEGN INCREASED
NO
READ THE NEXT COUNTER
YES
CONVERT COUNTER VALUE TO BCD
GET RDELET WORD WHICH CORRESPONDS
STORE RDELET WORD WITH COUNTERS
MOVE RDELET WORD PLUS ALL FOUR CORRESPONDING COUNTERS TO PRINT ARRA
IS THIS THE LAST COUNTER?
NO
YES
PRINT OUT COUNTERS AND RDELET WORDS
TRA 1,4


S/R TO INCREASE ONE OF FOUR COUNTERS FOR EACH RDELET WORD ENCOUNTERED
ACCUMULATOR CONTAINS THE NUMBER OF THE ONE OF FOUR COUNTERS TO BE INCREASED FOR EACH RDELET WORD
I1 CONTAINS THE ADDRESS OF THE CURRENT RDELET WORD
CTR
LOAD INDEX REGISTER 1 FROM I1
PLACE THE ACCUMULATOR IN INDEX REGISTER 2
DECREMENT OF THE THIRD ENGLISH COMPUTER WORD OF THE CURRENT SENTENCE CONTAINS THE RDELET WORD WHOSE COUNTER IS TO BE INCREASED
ADD A ONE TO THE APPROPRIATE COUNTER
CTRX
TRA 1,4